Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Information Theoretical Aspects of Complex Systems

Lecture 2.06

EEU45C09 / EEP55C09

Self Organising Technological Networks

CONNECT
Networks of the Future

Nicola Marchetti
nicola.marchetti@tcd.ie

Markov processes and entropy

- Some symbol sequences might be the direct result of a *Markov process*, where the probability for the next state (or symbol) is fully determined by the preceding state (or symbol)

- This means that the conditional probability distribution $P(\bullet \mid z_1,\ldots,z_{n-1})$ over the next symbol z_n, given a preceding sequence of symbols, converges already for $n = 2$, and Shannon entropy is then ΔS_2

$$s = \lim_{n\to\infty} \sum_{x_1\ldots x_{n-1}} p(x_1\ldots x_{n-1}) \sum_{x_n} p(x_n \mid x_1\ldots x_{n-1}) \log \frac{1}{p(x_n \mid x_1\ldots x_{n-1})} = \lim_{n\to\infty} \Delta S_n = \Delta S_\infty$$

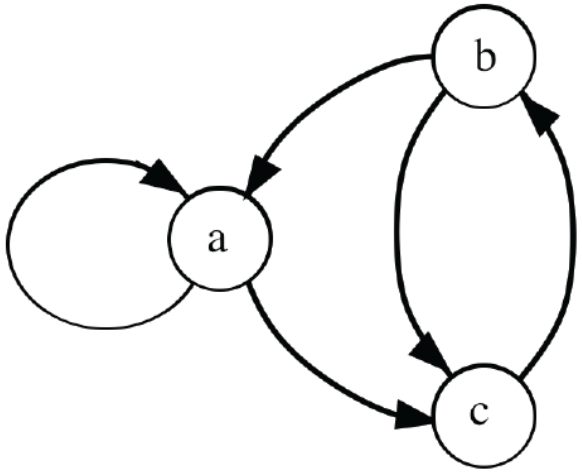
$$s = \sum_{z_1} p(z_1) \sum_{z_2} p(z_2 \mid z_1) \log \frac{1}{p(z_2 \mid z_1)}$$

Markov processes and entropy

- A Markov process can be described as a finite automaton with internal states z_i (with $i = 1,\ldots,m$, and z_i belonging to the alphabet Λ) corresponding to the symbols generated

- Process changes internal states according to transition probabilities, P_{ij}, denoting the probability to move from state i to state j

- Say the internal states (possible symbols) are a, b, c



Note: unless otherwise specified, we will assume that, when there is a choice for the transition from a state, then all possible transitions from that state are equally probable

Markov processes and entropy

- Hence

$$P_{aa} = P_{ac} = P_{ba} = P_{bc} = ½,\ P_{cb} = 1$$

$$P_{ab} = P_{ca} = P_{cc} = P_{bb} = 0$$

- To calculate the entropy of the process (or the symbol sequence), we need p(z), which is the probability distribution over the internal states

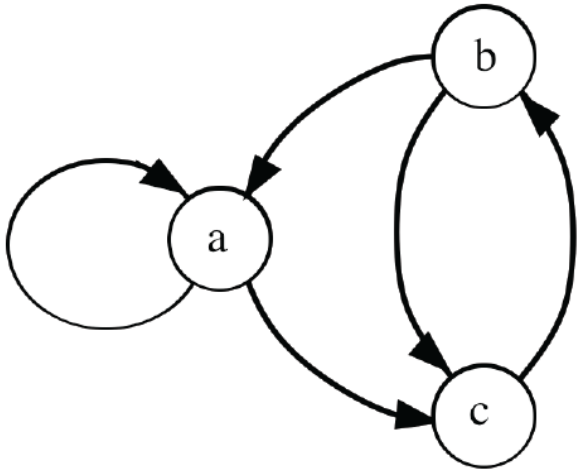
- Since the (stationary) probability p(z) to be in a certain state z must equal the sum over the probabilities of possible preceding states w weighted with the transition probabilities to state z, we can determine p(z) by the equations

(1) $$p(z) = \sum_w p(w) P_{wz} \quad , \text{for all } z \in \Lambda \qquad \Sigma_z\, p(z) = 1$$

Markov processes and entropy

- Note that the transition probabilities P_{zw} equal the conditional probabilities p(w|z) , which means the Shannon entropy s can be written as

$$s = \sum_z p(z) \sum_w p_{zw} \log \frac{1}{p_{zw}}$$



In this example the stationary distribution over the states is found by solving eqs. (1)

$$p(a) = \frac{1}{2} p(a) + \frac{1}{2} p(b),$$
$$p(b) = p(c),$$
$$p(c) = 1 - p(a) - p(b),$$

Markov processes and entropy

- One finds that p(a) = p(b) = p(c) = 1/3

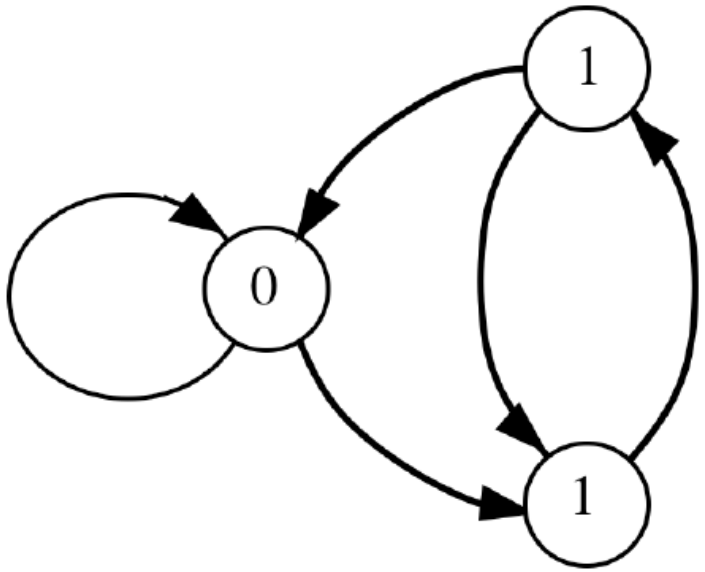
- There is an element of surprise only when the process is in state a or b, so the entropy turns out to be

$$s = p(a) \cdot I_a + p(b) + I_b + p(c) I_c = \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 0 = \frac{2}{3} \text{ bits}$$

Hidden Markov models and entropy

- In a *hidden Markov model*, one does not observe the states z of the process, but one observes some function of the state f(z)

- If in the example we just saw, the function f is given by f(a) = 0 , f(b) = f(c) = 1 , then we have a process that generates a sequence of 0 and 1 symbols

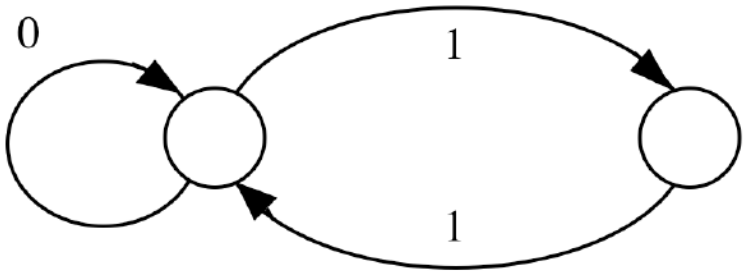


Hidden Markov models and entropy

- The process thus generates sequences of 0 and 1, with the restriction of having blocks of 1's (separated by 0's) always of even length

- We can illustrate the process in a different (but equivalent) form, by associating the symbols generated by the process with the *transitions* rather than with the internal *states*

- In that way we can get a more compact description



Note: while entropy is still s = 2/3 , but using the alternative description other information-theoretic quantities might change

Crystal

- Consider a periodic symbol sequence of 0 and 1

…0101010101010101010101010101…

- The density information measures the difference between the a priori uniform distribution $P_1^{(0)}$ and the observed single symbol distribution P_1

- $k_1 = 0$, since the probabilities are equal for the two symbols, p(0) = p(1) = 1/2. We can see that by using

$$k_1 = K[P_1^{(0)}; P_1] = \sum_{x_1} p(x_1) \log \frac{p(x_1)}{1/\nu} = \log \nu - S_1$$

- We have that

(2) $$p(0) = p(1) = \frac{1}{2}, \quad p(0|1) = p(1|0) = 1, \quad \text{and } p(0|0) = p(1|1) = 0$$

Crystal

- Using eqs.

$$K[P^{(0)};P] = \sum_{x_n} p(x_n \mid x_1 x_2 \ldots x_{n-1}) \log \frac{p(x_n \mid x_1 x_2 \ldots x_{n-1})}{p(x_n \mid x_2 \ldots x_{n-1})}$$

$$k_n = \sum_{x_1 \ldots x_{n-1}} p(x_1 \ldots x_{n-1}) K[P^{(0)};P]$$

- Plugging eq. (2) in the above equation, we can compute the correlation information from length 2 :

$$k_2 = \sum_{x_1} p(x_1) \sum_{x_2} p(x_2 \mid x_1) \log \frac{p(x_2 \mid x_1)}{p(x_2)} = \log 2 = 1 \text{ (bit)}$$

Crystal

- Since $k_2 = 1$ and 1 bit is also the total information (per symbol) of the system, then

$$1 = S_{\max} = k_{\text{corr}} + s$$
$$= \sum_{m=1}^{\infty} k_m + s = k_1 + k_2 + k_3 + \cdots + s = 0 + 1 + 0 + \cdots$$

- Hence, $k_m = 0$ for $m \neq 2$, and $s = 0$

- This is what we should expect, as there is no entropy in the system: as soon as we see one symbol (0 or 1), we also know the next symbol, and so on

Gas

- Consider a symbol sequence generated by a completely random process (e.g., coin tossing)

...1100001101000101100100111011101...

- The probability for next character to be 0 or 1 is 1/2, independently of how many preceding characters that we may observe

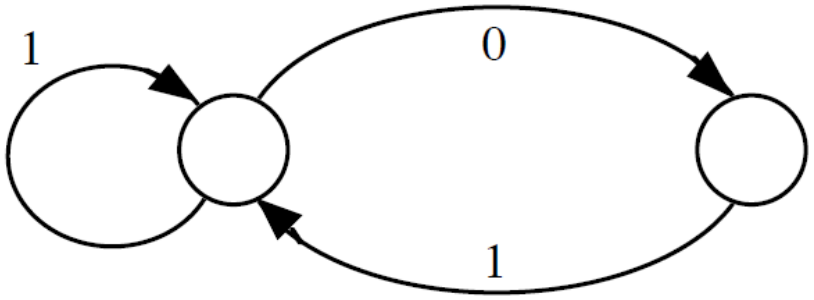
- The entropy of the conditional probability is therefore always maximal, s = $\log_2 2$ = 1 bit

- Therefore, there no contributions from redundancy, and k_{corr} = 0 . In fact,

$$1 = S_{\max} = k_{\text{corr}} + s = \sum_{m=1}^{\infty} k_m + s = 0 + 1$$

Finite automaton generating short correlations

- Consider a symbol sequence generated by the stochastic process described by this finite automaton



- This automaton generates an infinite sequence of 0's and 1's by following the arcs between the nodes in the graph

- When there are two arcs leaving a node, one is chosen randomly with equal probabilities for both choices

Finite automaton generating short correlations

- The above automaton generates a symbol sequence where 0's cannot appear in pairs, e.g.,

...1101011111011110101011101110110...

- To calculate the information-theoretic properties, we need to transform the characteristics of the automaton to a probabilistic description of the symbol sequence

- First, we calculate the densities of 0's and 1's

- Since the transition (arc) to the right node (R) always generates a 0 and is the only way a 0 can be generated, then the probability for being in R, p(R) equals the frequency of 0's, i.e., p(0) = p(R)

- Similarly, p(1) = p(L)

Finite automaton generating short correlations

- The probability for being in the left node p(L) must be equal to the probability that we were in this node last step and generated a 1 plus the probability that we were in the right node last step,

$$p(\mathrm{L}) = p(\mathrm{L})\frac{1}{2} + p(\mathrm{R})$$

$$p(\mathrm{L}) + p(\mathrm{R}) = 1$$

$$\Rightarrow \quad p(\mathrm{L}) = \frac{2}{3} \text{ and } p(\mathrm{R}) = \frac{1}{3}$$

- Or equivalently, p(0) = 1/3 and p(1) = 2/3

Finite automaton generating short correlations

- Then the density information is

$$k_1 = \sum_{x_1} p(x_1) \log \frac{p(x_1)}{1/2} = \frac{5}{3} - \log 3 \approx 0.0817$$

- If we observe one character, we know which node in the automaton is the starting point for generating the next character

$$p(0|1) = p(1|1) = 1/2,\ p(1|0) = 1,\ \text{and}\ p(0|0) = 0$$

- This means m = 2 . Thus, all other redundant information besides k_1 , is contained in correlations over block length 2

Finite automaton generating short correlations

- The correlation information for length two is then

$$k_2 = \sum_{x_1} p(x_1) \sum_{x_2} p(x_2 \mid x_1) \log \frac{p(x_2 \mid x_1)}{p(x_2)} = \log 3 - \frac{4}{3} \approx 0.2516$$

- The Shannon entropy s is the entropy (uncertainty) that remains when we are guessing the next character in the sequence, based on our knowledge on all preceding characters

- The preceding characters inform us on which node is used in generating the next character, and actually this information is in the last character alone

Finite automaton generating short correlations

- Formally, this can be expressed

$$p(1 \mid x_1 \ldots x_{n-1}1) = p(0 \mid x_1 \ldots x_{n-1}1) = \frac{1}{2}, \quad \text{for all possible } x_1 \ldots x_{n-1}1, \text{ and}$$

$$p(1 \mid x_1 \ldots x_{n-1}0) = 1, \quad \text{for all possible } x_1 \ldots x_{n-1}0 .$$

- Using the following equations

$$\left\langle S[P(\bullet \mid x_1 \ldots x_{n-1})\right\rangle = \sum_{x_1 \ldots x_{n-1}} p(x_1 \ldots x_{n-1}) \sum_{x_n} p(x_n \mid x_1 \ldots x_{n-1}) \log \frac{1}{p(x_n \mid x_1 \ldots x_{n-1})} =$$

$$= S_n - S_{n-1} = \Delta S_n .$$

$$s = \lim_{n \to \infty} \sum_{x_1 \ldots x_{n-1}} p(x_1 \ldots x_{n-1}) \sum_{x_n} p(x_n \mid x_1 \ldots x_{n-1}) \log \frac{1}{p(x_n \mid x_1 \ldots x_{n-1})} = \lim_{n \to \infty} \Delta S_n = \Delta S_\infty$$

Finite automaton generating short correlations

- We can calculate the entropy

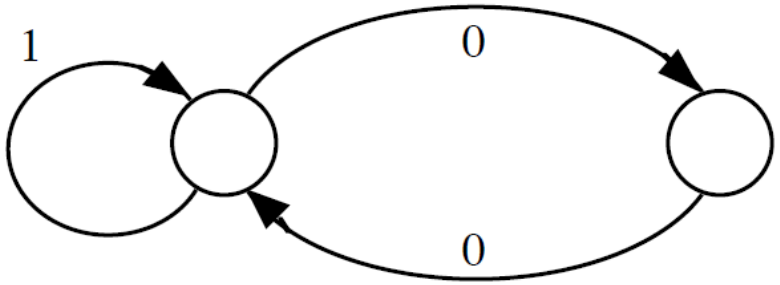
$$s = \lim_{m \to \infty} \Delta S_m = \Delta S_2 = \sum_{x_1} p(x_1) \sum_{x_2} p(x_2 \mid x_1) \log \frac{1}{p(x_2 \mid x_1)} =$$

$$= p(1) \cdot \log 2 + p(0) \cdot 0 = \frac{2}{3} \approx 0.6667$$

- This is what we should expect, since we have already said that there is no more correlation information than what was calculated in k_1 and k_2, and then the rest of the 1 bit of information per symbol must be the Shannon entropy s

Finite automaton generating long correlations

- The following finite automaton is similar to the one of the previous example, with the difference that the arc from R to L generates a 0



- This means that the automaton generates sequences where 1's are separated by an even number of 0's. This is a hidden Markov model

- Suppose that we shall guess the next character in the sequence, and that we may take into account a large (infinite) number of preceding characters, for example,

...0001000011100100000001100000?

Finite automaton generating long correlations

- Then it is sufficient to go back to the closest preceding ‘1’ and count how many 0's there are in between that symbol 1 and the present instant

- If there is an even number (including zero 0's), we are in L, and if there is an odd number, we are in R

- When we know which node we are in, we also have the best probability description of the next character

- Note: only if there are only 0's to the left, no matter how far we look, we will not be able to tell which is the node, but as the length of the preceding sequence tends to infinity the probability for this to happen tends to zero

Finite automaton generating long correlations

- Since the preceding sequence almost always determines (and corresponds to) the actual node, in the limit of infinite length, we can rewrite the entropy s as follows

$$s = \lim_{n\to\infty} \sum_{x_1 \ldots x_{n-1}} p(x_1 \ldots x_{n-1}) \sum p(x_n \,|x_1 \ldots x_{n-1}) \log \frac{1}{p(x_n \,|\, x_1 \ldots x_{n-1})} =$$

$$= p(\mathrm{L}) \sum_x p(x \,|\mathrm{L}) \log \frac{1}{p(x \,|\mathrm{L})} + p(\mathrm{R}) \sum_x p(x \,|\mathrm{R}) \log \frac{1}{p(x \,|\mathrm{R})} =$$

$$= \frac{2}{3} \cdot \log 2 + \frac{1}{3} \cdot 0 = \frac{2}{3} \approx 0.6667$$

- We have used the probabilities for the nodes L and R from the previous example, since they are the same

Finite automaton generating long correlations

- Expressed in this way, it is clear that the entropy of the symbol sequence comes from the random choice the automaton has to make in node L
- Node R does not generate any randomness or entropy
- We are therefore bound to get the same entropy as in the former example
- How about k_1 ? Given that

$$p(0) = p(R|L)p(L) + p(L|R)p(R) = \frac{1}{2} \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} = \frac{2}{3}$$

$$p(1) = p(L|L)p(L) = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}$$

- We will get the same value as before

Finite automaton generating long correlations

- When we get to the correlation information over longer blocks though, we get a difference

- In this case, we have correlation information in arbitrarily large blocks, since any number of consecutive 0's may occur, and then there is always more information to be gained by observing one more character

- To calculate the expressions for the different correlation information terms is now more complicated, and it is possible (but difficult) to prove that

$$k_{2n-1} = \frac{1}{3 \cdot 2^n}(9\log 3 - 14) \ , \ n = 1, 2, 3, \ldots , \text{and}$$

$$k_{2n} = \frac{1}{3 \cdot 2^{n-1}}(5 - 3\log 3), \ n = 1, 2, 3, \ldots$$

Finite automaton generating long correlations

- So, even if the last two discussed examples are identical in terms of entropy and redundancy, this second example *may be considered more complex, since the correlation information is spread out over larger distances*

- In the next lecture, we will see that such a difference may be used as one way to characterise the complexity of symbol sequences

Ackowledgement

- Kristian Lindgren, "Information Theory for Complex Systems", pages 24-30